Extending Landauer's Bound from Bit Erasure to Arbitrary Computation

David H. Wolpert[1, *]

[1] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
`http://davidwolpert.weebly.com`

(Dated: November 26, 2015)

Recent analyses have calculated the minimal thermodynamic work required to perform a computation π when two conditions hold: the output of π is independent of its input (e.g., as in bit erasure); we use a physical computer C to implement π that is specially tailored to the environment of C, i.e., to the precise distribution over C's inputs, $\mathcal{P}_0$. First I extend these analyses to calculate the work required even if the output of π depends on its input, and even if C is not used with the distribution $\mathcal{P}_0$ it was tailored for. Next I show that if C will be re-used, then the minimal work to run it depends only on the logical computation π, independent of the physical details of C. This establishes a formal identity between the thermodynamics of (re-usable) computers and theoretical computer science. I use this identity to prove that the minimal work required to compute a bit string σ on a "general purpose computer" rather than a special purpose one, i.e., on a universal Turing machine U, is $k_BT\ln(2)$[Kolmogorov complexity(σ) + log (Bernoulli measure of the set of strings that compute σ) + log(halting probability of U)]. I also prove that using C with a distribution over environments results in an unavoidable increase in the work required to run the computer, even if it is tailored to that distribution over environments. I end by using these results to relate the free energy flux incident on an organism / robot / biosphere to the maximal amount of computation that the organism / robot / biosphere can do per unit time.

There has been great interest for over a century in the relationship between thermodynamics and computation [2, 6, 11, 17, 18, 29–31, 38, 39, 46, 47, 49, 52–54]. A breakthrough was made with the argument of Landauer that at least $kT\ln[2]$ of work is required to run a 2-to-1 map like bit-erasure on any physical system [1–4, 12, 18, 26–28, 33, 35, 42, 45], a conclusion that is now being confirmed experimentally [5, 13, 24, 25, 40]. A related conclusion was that a 1-to-2 map can act as a *refrigerator* rather than a heater, *removing* heat from the environment [2–4, 26]. For example, this occurs in adiabatic demagnetization of an Ising spin system [26].

This early work leaves many issues unresolved however. In particular, say any output can be produced by our map, with varying probabilities, from any input. So the map is neither a pure heater nor a pure refrigerator. What is the minimal required work in this case?

More recently, there has been dramatic progress in our understanding of non-equilibrium statistical physics and its relation to information-processing [7, 9, 10, 12, 14–16, 20, 23, 34, 36, 37, 42–44, 48, 50]. Much of this recent literature has analyzed the minimal work required to drive a physical system's (fine-grained, microstate) dynamics during the interval from $t = 0$ to $t = 1$ in such a way that the dynamics of the macrostate is controlled by some desired Markov kernel π. In particular, there has been detailed analysis of the minimal work needed when there are only two macrostates, $v = 0$ and $v = 1$, and we require that both get mapped to the bin $v = 0$ [15, 34, 41]. By identifying the macrostates $v \in V$ as Information Bearing Degrees of Freedom (IBDF [4]) of an information-processing device like a digital computer, these analyses can be seen as elaborations of the analyses of Landauer et al. on the thermodynamics of bit erasure.

Many of the work-minimizing systems considered in this recent literature proceed in two stages. First, they physically change an initial, non-equilibrium distribution over microstates to the equilibrium distribution, $\rho^{eq}(w)$, in a quenching process. All information concerning the initial microstate is lost from the distribution over w by the end of this first stage. So in particular all information is lost about what the initial bin v_0 was. In addition, the Hamiltonian used in this quench is defined in terms of $\mathcal{P}_0$, the initial distribution over computer inputs. There is some unavoidable extra work if the computer is used with an initial distribution that differs from $\mathcal{P}_0$.

Next, in the second stage $\rho^{eq}(w)$ is transformed to an ending (non-equilibrium) distribution over w, with an associated distribution over the ending coarse-grained bin, v_1. However since all information about v_0 has been lost by the beginning of the second stage, v_0 cannot have any effect on the distribution over v_1 produced in the second stage. Accordingly, changing the distribution over inputs to one of these systems has no effect on the distribution over outputs. So although such a system can be used to implement a many-to-one map over the IBDF (i.e., the bins) in a digital computer, it cannot be used to implement any computational map whose output varies with its input.

In this paper I show how to implement any given conditional distribution π with minimal work, even if π maps different initial macrostates v_0 to different final macrostates v_1. I do this by connecting the original, **processor** system with macrostates $v \in V$ to a separate, initialized "memory system" that records v_0, and then evolve the joint system in such a way that the processor dynamics effectively samples $\pi(. \mid v_0)$. After this the memory is re-initialized (i.e., the stored copy of v_0 is erased), completing the cycle.

Like the systems considered in the literature, those considered here are implicitly optimized for some "prior" distribution over the inputs, $\mathscr{G}_0$. Here I go beyond the analyses in the literature by allowing the actual distribution over inputs, $\mathcal{P}_0(v)$, to differ from our assumed distribution, $\mathscr{G}_0$. When $\mathscr{G}_0 = \mathcal{P}_0$, the dynamics of the joint system is thermodynamically reversible. So the second law tells us that there is no alternative system that implements π with less work. However if $\mathcal{G}_0 \neq \mathcal{P}_0$ (i.e., the computer is used with a different user from the one they are optimized for) and π is not just a permutation over v, some of the work when the memory is reinitialized is unavoidably wasted. I then analyze the situation where there is a distribution over $\mathcal{P}_0$ (e.g., as occurs if the system is a computer that will be used with multiple users, or if it is an organism that will experience different environments) and $\mathcal{G}_0$ is optimized for that distribution, deriving how much extra work is needed due to uncertainty about who the user is.

I also show that *if the physical system used to run the computation will be re-used*, then the "internal entropies", giving the entropy internal to each coarse-grained bin, do not contribute to the minimal work. In such a scenario the specifics of the physical system implementing the computation — which are reflected in those internal entropies — are irrelevant. The work depends only on the computation π implemented by that system. (In previous analyses the computer was not re-used, so the internal entropies — and therefore physical details of the computer — were relevant.) This result establishes a formal identity between the thermodynamics of (re-usable) computers and theoretical computer science.

As an illustration, I use this identity to analyze the thermodynamics of a "general purpose computer" rather than a special purpose one, i.e., of a universal Turing machine U, where the macrostates are labelled by bit strings. In particular I prove that the work required to compute a particular bit string σ on U is $k_B T \ln(2)$ times the sum of the Kolmogorov complexity of σ, log of the Bernoulli measure of set of all strings that compute σ, and log of the Halting probability for U. Intuitively, by considering *all* input strings that result in σ, the second term quantifies "how many-to-one" U is, something that is not captured by the Kolmogorov complexity of σ.

I end by using these results to relate the free energy flux incident on an organism (robot, biosphere) to the maximal "rate of computation" implemented by that organism (resp., robot, biosphere).

I refer to the engineer who constructs the system as its "designer", and refer to the person who chooses its initial state as its "user". While the language of computation is used throughout this paper, the analysis applies to any dynamic process π over a coarse-grained partition of a fine-grained space, not just those processes conventionally viewed as computers. So for example, the analysis applies to the dynamics of biological organism reacting to its environment, if we coarse-grain that dynamics; the organism is the "computer", the dynamics is the "computation", the "designer" of the organism is natural selection, and the "user" initializing the organism is the environment.

Problem setup — I write $|X|$ for the number of elements x in any finite space X, and write the Shannon entropy of a distribution p over X as $S_p(X) = S(p) = -\sum_x p(x)\ln[p(x)]$, or even just $S(X)$ when p is implicit. I use similar notation for conditional entropy, etc. I also write the cross-entropy between two distributions p and q both defined over some space X as $C(p(X) \parallel q(X)) \equiv -\sum_x p(x)\ln[q(x)]$ or sometimes just $C(p \parallel q)$ for short [8, 32].

Let W be the space of all possible microstates of a system and $\mathcal{V}$ a partition of W, i.e., a coarse-graining of it into macrostates. For example, in a digital computer, $\mathcal{V}$ maps each microstate of the computer, $w \in W$, into the bit pattern in the computer's memory. I assume that the set of labels of the partition elements, V, contains "0". When convenient, I subscript a partition element with a time that the system state lies in that element, e.g., writing, v_0, v_1, etc.

The Hamiltonian over W at $t = 0$ is $H^{\varnothing}_{sys}$, with associated equilibrium (Boltzmann) distribution ρ^{eq}. For simplicity, I assume that $\forall v \in V$, at the two times $t = 0$ and $t = 1$, $Pr(w \mid v)$ is the same distribution, which I write as $q^v_{in}(w)$. (N.b., $q^v_{in}(w) = 0$ if $\mathcal{V}(w) \neq v$.) As in the analyses of computers in [2–4, 26], there is a "user" of the system who intervenes in its dynamics at or before $t = 0$, which results in the initial macrostate $v_0 \in V$ being set by sampling a **user distribution** $\mathcal{P}_0(v)$. As examples, $\mathcal{P}_0$ could model randomness in how a single user of a computer initializes the computer at $t = 0$, or randomness in how an environment of an organism perturbs the organism at $t = 0$. I write the (potentially non-equilibrium) unconditioned distribution over W at $t = 0$ as $\mathcal{P}_0(w) \equiv \sum_v \mathcal{P}_0(v) q^v_{in}(w)$.

The evolution of the microstates $w \in W$ during $t \in [0, 1)$ results in a conditional distribution over macrostates, $\pi(v_1 \mid v_0)$. Since they are set by the designer of the system, I take π and the distributions q^v_{in} to be fixed and known to that designer. However I allow the designer to be uncertain about what $\mathcal{P}_0$ is. As shorthand, I write $\mathcal{P}_1(v) \equiv \sum_v \mathcal{P}_0(v)\pi(v \mid v)$

I wish to focus on the component of the thermodynamic work that reflects computation, ignoring the component that reflects physical labor. This is guaranteed if the expected value of the Hamiltonian at $t = 0$ and $t = 1$ is the same, regardless of $\mathcal{P}_0$ and $\mathcal{P}_1$, since that means that the change in the expected value of the Hamiltonian is zero. Accordingly I assume that at both $t = 0$ and $t = 1$, the expected value of the Hamiltonian if the system is in state v then (i.e., $\sum_w q^v_{in}(w) H^{\varnothing}_{sys}(w)$) is a constant independent of v. I write that constant as $h^{\varnothing}_{sys}$. To simplify the analysis below, I also assume that $\mathbb{E}_{\rho^{eq}}[H^{\varnothing}_{sys}(w)] = h^{\varnothing}_{sys}$.

Overview of the system — The designer's goal is to modify the system considered in [15, 34, 41] into one which no longer loses the information of what the initial macrostate v_0 was as it evolves from $t = 0$ to $t = 1$. This can be done by coupling the system with an error-free **memory apparatus**, patterned after the measurement apparatus introduced in [34, 41, 42]. As in those studies, the "measurement" is a process that copies the macrostate to an initialized, external, symmetric memory with the value of v_0, and does so without changing v_0 (or even the initial microstate of the processor, w_0). Having set the value of such a memory, we can use its value later on, to govern the

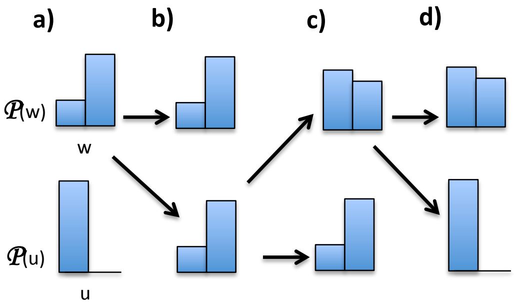

FIG. 1. Example of dynamics of the marginal distributions of a system with a binary coarse-graining, where the bins have the same size and both $q^v_{in}(w)$ and $Q^m(u)$ are uniform for all v, m. The top row shows the dynamics of the processor, and the bottom row shows the dynamics of the memory. Fig. (a) shows the $t = 0$ state, with the right bin of the processor more probable than the left bin, and the memory is in its initialized bin. Fig. (d) shows the $t = 1$ state, where the relative probabilities of the processor bins have changed according to π, and the memory has been returned to its initialized bin.

dynamics of w after the time when the the system distribution has relaxed to $\rho^{eq}(w)$, to ensure that v evolves according to π — even if under π the ending macrostate of the system depends on its initial state. Finally, to complete a cycle, the memory apparatus must be reinitialized.

I assume that the memory and system are both always in contact with a heat bath at temperature T. To be able to store a copy of any $v \in V$, the memory must have the same set of possible macrostates, V. I write the separate memory macrostates as $m \in V$, with associated microstates $u \in U$. (*A priori*, U need not have any relation to W.) For simplicity I assume that the conditional distribution of u given any m is the same distribution $Q^m_t(u)$ at both $t = 0$ and $t = 1$, and that there is a uniform equilibrium Hamiltonian $H^{\varnothing}_{mem}(u)$. In addition, I make the inductive hypothesis that the starting value of the memory is $m = 0$, with probability 1. The system dynamics comprises the following four steps (see Fig. 1):

I — First the memory apparatus copies the initial value $v = v_0$ into the memory, i.e., sets $m = v_0$. This step is done without any change to w, and so $\mathcal{P}_0(w)$ is unchanged. Since the copy is error-free and the memory is symmetric, this step does not require thermodynamic work [34].

II — Next a Quench-then-Relax procedure (QR) like the one described in [15, 34] is run on the distribution over w, $q^{v_0}_{in}(w)$. In such a QR, first we replace $H^{\varnothing}_{sys}$ with a **quenching Hamiltonian** chosen such that $q^{v_0}_{in}$ is an equilibrium distribution for a Hamiltonian specified by the memory system macrostate:

$$H^m_{in}(w) \equiv -kT \ln[q^{v_0}_{in}(w)] \tag{1}$$

(While w is unchanged in this adiabatic quench, and therefore so is the distribution over W, in general work is required if $H^m_{in} \neq H^{\varnothing}_{sys}$.) Next we isothermally and quasi-statically relax H^m_{in} back to $H^{\varnothing}_{sys}$, thereby changing $q^{v_0}_{in}(w)$ to $\rho^{eq}(w)$. (See also [41].)

III — Next we use the fact that $m = v_0$ to run a QR over W in reverse, with the quenching Hamiltonian

$$H^m_{out}(w) \equiv -kT \ln[q^{v_0}_{out}(w)] \tag{2}$$

where $q^{v_0}_{out}(w) \equiv \sum_{v_1} q^{v_1}_{in}(w)\pi(v_1 \mid v_0)$. This reverse QR begins by isothermally and quasi-statically sending $H^{\varnothing}_{sys}$ to H^m_{out}. After that H^m_{out} is replaced by $H^{\varnothing}_{sys}$ in a "reverse quench", with no change to w. As in step (II), there is no change to m in step (III).

IV — Finally, as described in detail below, we reset m to 0. This ensures we can rerun the system, and also guarantees the inductive hypothesis.

Since the system samples $\pi(v_1 \mid v_0)$ in step (III), these four steps implement the map π even if π's output depends on its input, and no matter what $\mathcal{P}_0$ is. (The whole reason for storing v_0 in m was to allow this step III.)

Moreover, the expected work expended in the first three steps is given (with abuse of notation) by the conditional entropy,

$$-kT\Big(S_{\mathcal{P}}(V_1 \mid V_0) + \sum_v S(q^v_{in})\Big[\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big]\Big) \tag{3}$$

(See Supplementary Material (SM) for proof.)

Resetting the memory — We implement step (IV) by first running a QR on the distribution over U (not W), and then running a reverse QR, one that ends with $m = 0$ no matter what the initial value of m was.

In detail, suppose that the designer of the system guesses that the distribution over the initial values of the macrostates is $\mathscr{G}_0(v_0)$ — which in general need not equal $\mathcal{P}_0$. This distribution would be the prior probability over the values of m if G_0 equalled $\mathcal{P}_0$, since m is a copy of v_0. The associated likelihood of v_1 given m is $\mathscr{G}(v_1 \mid m) = \pi(v_1 \mid v_0 = m)$. So the posterior probability of m given v_1, $\mathscr{G}(m \mid v_1)$, is proportional to $\mathscr{G}_0(m)\pi(v_1 \mid m)$. This gives the (guessed) posterior probability over memory microstates, which we can write as

$$\mathscr{G}(u \mid v_1) = \sum_m \mathscr{G}(m \mid v_1) Q^m(u) \tag{4}$$

with some abuse of notation. In contrast, the actual posterior distribution $\mathcal{P}(m \mid v_1)$ is given by the actual prior $\mathcal{P}_0$, and gives a posterior distribution

$$\mathcal{P}(u \mid v_1) = \sum_m \mathcal{P}(m \mid v_1) Q^m(u) \tag{5}$$

The premise of this paper is that to reset the memory the computer first runs a QR using the quenching Hamiltonian

$$H^{v_1}_{mem}(u) \equiv -kT \ln \mathscr{G}(u \mid v_1) \tag{6}$$

to drive the distribution over u to $\rho^{eq}_{mem}(u)$, since this would relax the memory using minimal work if the guessed prior $\mathscr{G}_0$ equaled the actual one, $\mathcal{P}_0$. (Intuitively, v_1 is a "noisy measurement" of m that is used to set this quenching Hamiltonian, and we are running the same process as in step II, just with the roles of the memory and processor reversed.) Next we run a reverse QR, taking $\rho^{eq}_{mem}(u)$, the uniform distribution over all U, to the distribution that is uniform over $m = 0$, zero elsewhere. This completes the resetting of the memory macrostate.

Averaging the work required in this resetting of the memory, and adding it to the expression in Eq. (3), gives the minimal expected work for running π:

$$\Omega_{\mathscr{G}_0,\mathcal{P}_0} \equiv -kT\Big(\sum_{v_0,v_1} \mathcal{P}(v_0, v_1) \ln\left[\mathscr{G}(v_0 \mid v_1)\right] + S_{\mathcal{P}}(V_1 \mid V_0) \;+\; \sum_v S(q^v_{in})\Big[\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big]\Big) \tag{7}$$

(See SM for proof.) Since $\mathscr{G}(v_1 \mid v_0) = \pi(v_1 \mid v_0) = \mathcal{P}(v_1 \mid v_0)$, we can use Bayes' theorem to rewrite Eq. (7) as

$$kT\Big(C[\mathcal{P}(V_0) \,||\, \mathscr{G}(V_0))] - C[\mathcal{P}(V_1) \,||\, \mathscr{G}(V_1))] + \sum_v S(q^v_{in})\Big[\mathcal{P}_0(v) - \mathcal{P}_1(v)\Big]\Big) \tag{8}$$

(assuming all distributions over V have the same support, so that we don't divide by zero when using Bayes' theorem).

So if $\mathscr{G}_0 = \mathcal{P}_0$, or alternatively π is an invertible function over V, $\Omega_{\mathscr{G}_0,\mathcal{P}_0} = kT[S_{\mathcal{P}_0}(W) - S_{\mathcal{P}_1}(W)]$. This quantity is sometimes called "generalized Landauer cost". Note that for a fixed $P(w_0, w_1)$, it is independent of the partition.

Multiple cycles of a computer — Sometimes we will want to use an (IID) **calculator** computer, in which we IID sample $\mathcal{P}_0$ at the end of each iteration, over-writing v_1, before running π again. In such calculators, after step (IV) above, the value v_1 is copied to an external system via an additional memory apparatus (e.g., in order to drive some physical actuator). Then a different external system (e.g., a sensor) forms a sample $v'_0 \sim \mathcal{P}_0$, and v_1 gets replaced by v'_0. Only after these two new steps have we completed a full cycle. At this point we can run another cycle, to apply π again — but starting from v'_0 rather than v_1.

In the SM it is shown that for an "extended" calculator computer, where π is iterated N times and only then is v copied to an external system and v'_0 copied in, the total work expended is at least

$$kT\Big(C[\mathcal{P}(V_0) \,||\, \mathscr{G}(V_0)] - C[\mathcal{P}(V_N) \,||\, \mathscr{G}(V_N)]\Big) \tag{9}$$

Note that the expected work of a calculator has no dependence on the values $S(q^v_{in})$; in calculators the work depends only the logical map that π implements over V, independent of the physical system that implements that map. So there is a formal identity between the thermodynamics of (calculator) computers and computer science.

As an example, fix a prefix-free universal Turing machine U [19, 21, 51]. Identify the macrovariable $v \in V$ of the physical system implementing U with the instantaneous description (ID) of U, so that π gives the dynamics over those IDs, i.e., it is the dynamical law implementing the Turing machine. I will say that an instantaneous description (ID) of U is a **starting** ID if it specifies that the machine U is in its initial state with an input string for which U halts. Also define I^{σ} as the set of all starting IDs that halt with output σ, and $K_U(\sigma)$ as the Kolmogorov complexity of σ. Finally, for any starting ID α, define $\ell(\alpha)$ as the length of the input string of α.

Perhaps the most natural prior probability distribution over IDs is the (normalized) "universal prior probability", i.e., $\mathscr{G}_0(v) = 2^{-\ell(v)}/\lambda$ if v is a starting ID and $\mathscr{G}_0(v) = 0$ otherwise, where λ is U's halting probability. It is shown in the SM that under the simpler of two natural definitions of how to use a TM U as a "calculator computer", the minimal work (over all possible $\mathcal{P}_0$(needed to compute σ is

$$kT \ln(2)\Big(K_U(\sigma) + \log[\mathscr{G}_0(I^{\sigma})] + \log\lambda\Big) \tag{10}$$

So the greater the gap between the log-probability that a randomly chosen program computes σ and the log-probability of the most likely such program, the greater the work to compute σ. Intuitively, running U on I^{σ} executes a many-to-one map in the Landauer sense, taking many starting IDs to the same ending ID. The gap between $\log[\mathscr{G}_0(I^{\sigma})]$ and $\min_{v_0 \in I^{\sigma}} \log[\mathscr{G}_0(v_0)] = K_U(\sigma) + \log\lambda$ quantifying "how many-to-one" that map is. (Similar results hold for other choices of space of logical variables V, machine π and / or prior $\mathscr{G}_0$.)

As an aside, by Levin's coding theorem [51], $K_U(\sigma) + \log[\mathscr{G}_0(I^{\sigma})]$ is bounded by a constant that depends only on U, and is independent of σ. So for any U, there is a σ-independent upper bound on the minimal amount of work needed for U to compute σ.

Multiple users — Often rather than a single user of a calculator computer there will be a distribution over users, $Pr(\mathcal{P}_0)$. To analyze this situation, use Eq. (7) to write

$$\langle \Omega_{\mathscr{G}_0,\mathcal{P}_0} \rangle = \Omega_{\mathscr{G}_0,\langle \mathcal{P}_0 \rangle} \tag{11}$$

(where $\langle . \rangle$ indicates an average according to $Pr(.)$). Applying this equality to Eq. (9), and using the facts that Kullbach Leibler (KL) divergence is non-increasing in t and is minimized (at zero) when its arguments are equal [8], we see that the $\mathscr{G}_0$ that minimizes expected work is $\langle \mathcal{P}_0 \rangle$. The associated expected work is $S_{\langle \mathcal{P} \rangle}(\mathcal{V}_0) - S_{\langle \mathcal{P} \rangle}(\mathcal{V}_1)$.

The expected work would instead be $\langle S_{\mathcal{P}}(\mathcal{V}_0) - S_{\mathcal{P}}(\mathcal{V}_1) \rangle$ if we could somehow re-optimize $\mathscr{G}_0$ for each $\mathcal{P}_0$. So the difference between those two values of expected work can be viewed as the minimal penalty we must pay due to uncertainty about who the user is. This penalty can be re-expressed as the drop from $t = 0$ to $t = 1$ in the **entropic variance**,

$$\langle \mathcal{P} \ln[\mathcal{P}] \rangle - \langle \mathcal{P} \rangle \ln[\langle \mathcal{P} \rangle] \tag{12}$$

i.e., it is the growth from $t = 0$ to $t = 1$ in certainty about $\mathcal{P}$.

Entropic variance is non-negative and non-increasing.[1] So the work penalty that arises due to growth in certainty about $\mathcal{P}$ is always non-negative. This is true even if the minimal work required to implement the underlying computation is negative.

Implications for biology. — Any work expended on the processor must first be acquired as free energy from the processor's environment. However in many situations there is a limit on the flux of free energy through a processor's immediate environment. Combined with the analysis above, such limits provide upper bounds on the "rate of (potentially noisy) computation" that can be achieved by a biological organism in that environment. In particular, since the minimal work required to do a computation increases if $\mathscr{G}_0 \neq \mathcal{P}_0$, using the same biological organism in a new environment, differing from the one it is tailored for, will in general result in extra required work.

As an example, these results bound the rate of computation of a human brain. Given the fitness cost of such computation (the brain uses $\sim 20\%$ of the calories used by the human body), this bound contributes to the natural selective pressures on humans, in the limit that operational inefficiencies of the brain have already been minimized. In other words, these bounds suggest that natural selection imposes a tradeoff between the fitness quality of a brain's decisions, and how much computation is required to make those decisions. In this regard, it is interesting to note that the brain is famously noisy — and as discussed above, noise in computation reduces the total thermodynamic work required.

As a second example, the rate of solar free energy incident upon the earth provides an upper bound on the rate of computation that can be achieved by the biosphere. (This bound holds for any choice for the partition of the biosphere's fine-grained space into macrostates such that the dynamics over those macrostates executes π.) In particular it provides an upper bound on the rate of computation that can be achieved by human civilization, if we remain on the surface of the earth, and only use sunlight to power our computation.

Acknowledgements — I would like to thank Josh Grochow, Cris Moore, Daniel Polani, Simon DeDeo, Damian Sowinski, Eric Libby, and especially Sankaran Ramakrishnan for many stimulating discussions, and the Santa Fe Institute for helping to support this research. This paper was made possible through the support of Grant No. TWCF0079/AB47 from the Templeton World Charity Foundation and Grant No. FQXi-RHl3-1349 from the FQXi foundation. The opinions expressed in this paper are those of the author and do not necessarily reflect the view of Templeton World Charity Foundation.

[1] Resp., since entropy is a concave function of distributions, and since entropic variance is the average (over $\mathcal{P}_0$'s) of the KL divergence between $\mathcal{P}$ and $\langle \mathcal{P} \rangle$.

[*] Massachusetts Insitutute of Technology; Arizona State University

[1] Charles H Bennett, *Logical reversibility of computation*, IBM journal of Research and Development **17** (1973), no. 6, 525–532.

[2] ———, *The thermodynamics of computation—a review*, International Journal of Theoretical Physics **21** (1982), no. 12, 905–940.

[3] ———, *Time/space trade-offs for reversible computation*, SIAM Journal on Computing **18** (1989), no. 4, 766–776.

[4] ———, *Notes on landauer's principle, reversible computation, and maxwell's demon*, Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics **34** (2003), no. 3, 501–510.

[5] Antoine Bérut, Artak Arakelyan, Artyom Petrosyan, Sergio Ciliberto, Raoul Dillenschneider, and Eric Lutz, *Experimental verification of landauer/'s principle linking information and thermodynamics*, Nature **483** (2012), no. 7388, 187–189.

[6] L. Brillouin, *Science and information theory*, Academic Press, 1962.

[7] Farid Chejne Janna, Fadl Moukalled, and Carlos Andrés Gómez, *A simple derivation of crooks relation*, International Journal of Thermodynamics **16** (2013), no. 3, 97–101.

[8] T. Cover and J. Thomas, *Elements of information theory*, Wiley-Interscience, New York, 1991.

[9] Gavin E Crooks, *Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems*, Journal of Statistical Physics **90** (1998), no. 5-6, 1481–1487.

[10] ———, *Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences*, Physical Review E **60** (1999), no. 3, 2721.

[11] Lidia Del Rio, Johan Åberg, Renato Renner, Oscar Dahlsten, and Vlatko Vedral, *The thermodynamic meaning of negative entropy*, Nature **474** (2011), no. 7349, 61–63.

[12] Raoul Dillenschneider and Eric Lutz, *Comment on"minimal energy cost for thermodynamic information processing: Measurement and information erasure"*, Physical review letters **104** (2010), no. 19, 198903.

[13] Jörn Dunkel, *Thermodynamics: Engines and demons*, Nature Physics **10** (2014), no. 6, 409–410.

[14] Massimiliano Esposito and Christian Van den Broeck, *Three faces of the second law. i. master equation formulation*, Physical Review E **82** (2010), no. 1, 011143.

[15] ———, *Second law and landauer principle far from equilibrium*, EPL (Europhysics Letters) **95** (2011), no. 4, 40004.

[16] Philippe Faist, Frédéric Dupuis, Jonathan Oppenheim, and Renato Renner, *A quantitative landauer's principle*, arXiv preprint arXiv:1211.1037 (2012).

[17] Edward Fredkin, *An informational process based on reversible universal cellular automata*, Physica D: Nonlinear Phenomena **45** (1990), no. 1, 254–270.

[18] Edward Fredkin and Tommaso Toffoli, *Conservative logic*, Springer, 2002.

[19] Peter Grunwald and Paul Vitányi, *Shannon information and kolmogorov complexity*, arXiv preprint cs/0410002 (2004).

[20] H-H Hasegawa, J Ishikawa, K Takara, and DJ Driebe, *Generalization of the second law for a nonequilibrium initial state*, Physics Letters A **374** (2010), no. 8, 1001–1004.

[21] John E Hopcroft, Rajeev Motwani, and Ullman Rotwani, *Jd: Introduction to automata theory, languages and computability*, 2000.

[22] Jordan M Horowitz and Juan MR Parrondo, *Designing optimal discrete-feedback thermodynamic engines*, New Journal of Physics **13** (2011), no. 12, 123019.

[23] Christopher Jarzynski, *Nonequilibrium equality for free energy differences*, Physical Review Letters **78** (1997), no. 14, 2690.

[24] Yonggun Jun, Momčilo Gavrilov, and John Bechhoefer, *High-precision test of landauer's principle in a feedback trap*, Physical review letters **113** (2014), no. 19, 190601.

[25] JV Koski, VF Maisi, JP Pekola, and DV Averin, *Experimental realization of a szilard engine with a single electron*, arXiv preprint arXiv:1402.5907 (2014).

[26] Rolf Landauer, *Irreversibility and heat generation in the computing process*, IBM journal of research and development **5** (1961), no. 3, 183–191.

[27] ______, *Minimal energy requirements in communication*, Science **272** (1996), no. 5270, 1914–1918.

[28] ______, *The physical nature of information*, Physics letters A **217** (1996), no. 4, 188–193.

[29] Harvey S Leff and Andrew F Rex, *Maxwell's demon: entropy, information, computing*, Princeton University Press, 2014.

[30] Seth Lloyd, *Use of mutual information to decrease entropy: Implications for the second law of thermodynamics*, Physical Review A **39** (1989), no. 10, 5378.

[31] ______, *Ultimate physical limits to computation*, Nature **406** (2000), no. 6799, 1047–1054.

[32] D.J.C. Mackay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[33] O.J.E. Maroney, *Generalizing landauer's principle*, Physical Review E **79** (2009), no. 3, 031105.

[34] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa, *Thermodynamics of information*, Nature Physics **11** (2015), no. 2, 131–139.

[35] Martin B Plenio and Vincenzo Vitelli, *The physics of forgetting: Landauer's erasure principle and information theory*, Contemporary Physics **42** (2001), no. 1, 25–60.

[36] Blake S Pollard, *A second law for open markov processes*, arXiv preprint arXiv:1410.6531 (2014).

[37] Mikhail Prokopenko and Itai Einav, *Information thermodynamics of near-equilibrium computation*, Physical Review E **91** (2015), no. 6, 062143.

[38] Mikhail Prokopenko and Joseph T Lizier, *Transfer entropy and transient limits of computation*, Nature Scientific reports **4** (2014).

[39] Mikhail Prokopenko, Joseph T Lizier, and Don C Price, *On thermodynamic interpretation of transfer entropy*, Entropy **15** (2013), no. 2, 524–543.

[40] É Roldán, Ignacio A Martinez, Juan MR Parrondo, and Dmitri Petrov, *Universal features in the energetics of symmetry breaking*, Nature Physics (2014).

[41] Takahiro Sagawa, *Thermodynamic and logical reversibilities revisited*, Journal of Statistical Mechanics: Theory and Experiment **2014** (2014), no. 3, P03025.

[42] Takahiro Sagawa and Masahito Ueda, *Minimal energy cost for thermodynamic information processing: measurement and information erasure*, Physical review letters **102** (2009), no. 25, 250602.

[43] ______, *Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics*, Physical review letters **109** (2012), no. 18, 180602.

[44] Udo Seifert, *Stochastic thermodynamics, fluctuation theorems and molecular machines*, Reports on Progress in Physics **75** (2012), no. 12, 126001.

[45] Kousuke Shizume, *Heat generation required by information erasure*, Physical Review E **52** (1995), no. 4, 3495.

[46] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks, *Thermodynamics of prediction*, Physical review letters **109** (2012), no. 12, 120604.

[47] Leo Szilard, *On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings*, Behavioral Science **9** (1964), no. 4, 301–310.

[48] K Takara, H-H Hasegawa, and DJ Driebe, *Generalization of the second law for a transition between nonequilibrium states*, Physics Letters A **375** (2010), no. 2, 88–92.

[49] Tommaso Toffoli and Norman H Margolus, *Invertible cellular automata: A review*, Physica D: Nonlinear Phenomena **45** (1990), no. 1, 229–253.

[50] Hugo Touchette and Seth Lloyd, *Information-theoretic approach to the study of control systems*, Physica A: Statistical Mechanics and its Applications **331** (2004), no. 1, 140–172.

[51] M. Li and Vitanyi P., *An introduction to kolmogorov complexity and its applications*, Springer, 2008.

[52] Karoline Wiesner, Mile Gu, Elisabeth Rieper, and Vlatko Vedral, *Information-theoretic lower bound on energy cost of stochastic computation*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science **468** (2012), no. 2148, 4058–4066.

[53] W. H. Zurek, *Algorithmic randomness and physical entropy*, Phys. Rev. A **40** (1989), 4731–4751.

[54] Wojciech H Zurek, *Thermodynamic cost of computation, algorithmic complexity and the information metric*, Nature **341** (1989), no. 6238, 119–124.

DERIVATION OF EQ. 3 OF MAIN TEXT

In this section I evaluate the expected work required to implement the first three steps of the system for initial distribution $\mathcal{P}_0$. To do this, it will be convenient to calculate the expected work to perform those steps conditioned on a particular v_0, and then average over all v_0 according to $\mathcal{P}_0(v_0)$.

As in [34], I assume that after step (I) the interaction Hamiltonian between W and U is negligible. Also as in that work, I assume that the quench step at the beginning of step II is an instantaneous change to the energy of every w, $\Delta E(w)$. This process does not actually change w (such changes are associated with transfer of heat). Since the quenching Hamiltonian depends on the value of m (which due to step (I) depends on v_0), the value of $\Delta E(w)$ for each w also depends on m. That change in the energy of w is identified as the work done on the system in the quench step when it starts (and stays) in that state w.

Now due to the fact that step (I) did not change w, at the beginning of the quench step the posterior probability of w given a current value m is $q^m_{in}(w)$. Therefore the expected work done in this quench step conditioned on a particular value m is $\sum_w q^m_{in}(w)[H^m_{in}(w) - H^{\varnothing}_{sys}(w)]$. As shorthand, define S^0_{sys} as the Shannon entropy over W for the Boltzmann distribution with temperature T and Hamiltonian $H^{\varnothing}_{sys}$. Then conditioned on a value v_0 at the beginning of step (I), the work to perform the entire QR in step (II) is

$$\sum_w q^m_{in}(w)[H^m_{in}(w) - H^{\varnothing}_{sys}(w)] + \mathcal{F}(H^{\varnothing}_{sys}) - \mathcal{F}(H^m_{in}) \quad (13)$$

where $m = v_0$ and

$$\mathcal{F}(H^{\varnothing}_{sys}) = h^{\varnothing}_{sys} - kTS^0_{sys} \quad (14)$$

is the equilibrium free energy of $H^{\varnothing}_{sys}$ at temperature T.

By definition of H^m_{in}, $\mathcal{F}(H^m_{in}) = 0$. So the expression in Eq. (13) just equals $kT\Big[S(q^m_{in}) - S^0_{sys}\Big]$. Note that this amount of work is negative, since work is extracted by sending q^m_{in} to the equilibrium distribution for $H^{\varnothing}_{sys}$.

Similarly, to implement step (III) requires work of at least $kT\Big[S^0_{sys} - S(q^m_{out})\Big]$.[2] Now for any distribution $Pr(w)$, with some abuse of notation, we can write $S_{Pr}(v \mid w) = 0$, since w sets v uniquely. Therefore

$$\begin{aligned} S_{Pr}(w) &= S_{Pr}(v \mid w) + S_{Pr}(w) \\ &= S_{Pr}(v, w) \\ &= S_{Pr}(v) + S_{Pr}(w \mid v) \end{aligned} \quad (15)$$

So if we write the Shannon entropy of the distribution over values v_1 conditioned on a particular value of v_0 as

$$S_\pi(V_1 \mid v_0) \equiv -\sum_{v_1} \pi(v_1 \mid v_0) \ln[\pi(v_1 \mid v_0)] \quad (16)$$

then we can write

$$\begin{aligned} S(q^{v_0}_{out}) &= S_\pi(V_1 \mid v_0) - \sum_{w_1, v_1} \pi(v_1 \mid v_0) q^{v_1}_{in}(w_1) \ln(q^{v_1}_{in}(w_1)) \\ &= S_\pi(V_1 \mid v_0) + \sum_{v_1} \pi(v_1 \mid v_0) S(q^{v_1}_{in}) \end{aligned} \quad (17)$$

Accordingly, the total amount of work in the first three steps, conditioned on a value v_0, is

$$\begin{aligned} &kT\Big[S(q^{v_0}_{in}) - S(q^{v_0}_{out})\Big] \\ &\quad = kT\Big[S(q^{v_0}_{in}) - S_\pi(V_1 \mid v_0) - \sum_{v_1} \pi(v_1 \mid v_0) S(q^{v_1}_{in})\Big] \end{aligned} \quad (18)$$

Combining and averaging under $\mathcal{P}_0(v_0)$, the expected work required to complete the first three steps is

$$-kT\Big[S_\pi(V_1 \mid \mathcal{V}_0) + \sum_v S(q^v_{in})\Big(\mathcal{P}_1(v) - \mathcal{P}_0(v)\Big)\Big] \quad (19)$$

(The analogous expression in much of the literature has $S_\pi(V_1)$ instead of $S_\pi(V_1 \mid V_0)$; the difference is due to the requirement that π govern the coarse-grained dynamics even if its output depends on its input, a requirement that means that we must measure the value v_0.)

[2] In steps II and III the usual convention was followed by quasi-statically sending H^m_{in} to $H^{\varnothing}_{sys}$ and then sending $H^{\varnothing}_{sys}$ to $H^{v_0}_{out}$. The same total work would arise if we instead quasi-statically send H^m_{in} to $H^{v_0}_{out}$ directly.

DERIVATION OF EQ. 7 OF THE MAIN TEXT

The QR in resetting the memory is run at $t = 1$, using $H^{v_1}_{mem}(u)$. It does not change w_1, just as measurement of v_0 did not change w_0. Accordingly, the minimal amount of work in this QR is

$$\sum_u \mathcal{P}(u \mid v_1)[H^{v_1}_{mem}(u) - H^{\varnothing}_{mem}(u)] + \mathcal{F}(H^{\varnothing}_{mem})$$
$$= kT\Big(- \sum_u \mathcal{P}(u \mid v_1) \ln \Big[\mathscr{G}(u \mid v_1)\Big] - \ln |V|\Big) \qquad (20)$$

This is true whether or not $\mathscr{G}_0 = \mathcal{P}_0$. Note though that due to the fact that $H^{v_1}_{mem}$ is defined in terms of $\mathscr{G}_0(u \mid v_1)$ not $\mathcal{P}_0(u \mid v_1)$, if both $\mathscr{G}_0 \neq \mathcal{P}_0$ and π is not an invertible deterministic map, then the actual posterior $\mathcal{P}(u \mid v_1)$ is not the equilibrium distribution for $H^{v_1}_{mem}$. This means that immediately after the quenching process, as the Hamiltonian over U begins to quasi-statically relax, the distribution over U will first settle, in a thermodynamically irreversible process, to the equilibrium distribution for $H^{v_1}_{mem}$. No work is involved in that irreversible process. However if we had instead chosen $\mathscr{G}_0 = \mathcal{P}_0$, the expression in Eq. (20) would have been less, i.e., less work would have been required, since no such irreversible process would have occurred.

To complete resetting the memory we now run a reverse QR that takes u from the uniform distribution over all U to the distribution $Q^0(u)$, whose support is restricted to u's such that $m = 0$. This means that for the given value of v_1, the total work required to reset m to 0, including the contribution evaluated in Eq. (20), is

$$-kT\Big(\sum_u \mathcal{P}(u \mid v_1) \ln \Big[\mathscr{G}_0(u \mid v_1)\Big] \; + \; S(Q^0)\Big) \qquad (21)$$

Multiply and divide the argument of the logarithm in the summand by $\mathcal{P}(u \mid v_1)$. Next use the same kind of decomposition as in Eq. (15), and then use the chain-rule for KL divergence. This transforms our expression into

$$- kT\Big(\sum_{v_0} \mathcal{P}(v_0 \mid v_1) \ln \Big[\mathscr{G}_0(v_0 \mid v_1)\Big] - \sum_{v_0} \mathcal{P}(v_0 \mid v_1) S(Q^{v_0}) + S(Q^0)\Big) \qquad (22)$$

Averaging this according to $\mathcal{P}_1(v_1)$ gives

$$- kT\Big(\sum_{v_0, v_1} \mathcal{P}(v_0, v_1) \ln \Big[\mathscr{G}_0(v_0 \mid v_1)\Big] - \sum_{v_0} \mathcal{P}(v_0) S(Q^{v_0}) + S(Q^0)\Big) \qquad (23)$$

Note though that we assumed that the states of the memory are symmetric. (This is why there is no expected work in step (I).) So $S(Q^v)$ is independent of v, and Eq. (23) reduces to

$$- kT \sum_{v_0, v_1} \mathcal{P}(v_0, v_1) \ln \Big[\mathscr{G}_0(v_0 \mid v_1)\Big] \qquad (24)$$

Adding Eq. (24) to Eq. (3) of the main text gives Eq. (7) of the main text, as claimed.

DERIVATION OF EQ. 9 OF MAIN TEXT

Since no work is required in the new step where we measure v_1, the total work in an iteration is given by adding Eq. (8) to the additional average work required to map $v = v_1$ to $v = v'_0$. Since both the values v_1 and v'_0 exist outside of W, they can be used to specify the two quenching Hamiltonians that implement this map. So the additional average work is $kT \sum_{v,v'} \Big[S(q^{v}_{in})\mathcal{P}_1(v) - S(q^{v'}_{in})\mathcal{P}_0(v')\Big]$. Generalizing this reasoning gives

$$kT\Big(C[\mathcal{P}(V_0) \,||\, \mathscr{G}(V_0)] - C[\mathcal{P}(V_N) \,||\, \mathscr{G}(V_N)]\Big) \qquad (25)$$

as claimed.

Note that this result requires the computer to contain an integer-valued clock, whose state t increases by 1 at each iteration. This clock is needed so that the appropriate posterior $\mathscr{G}(m_t \mid v_{t+1})$ can be used to set $H^{v_{t+1}}_{mem}(u_t)$ at iteration t. Note that such a clock can be implemented without any work, since its dynamics is logically reversible. Given such a clock, the cross-entropies and internal entropies over iterations $t \in 2, \ldots, N - 1$ cancel out.

DERIVATION OF EQ. 10 OF MAIN TEXT

We are ultimately interested in the map from U's input tape to its output tape. In addition, U is a prefix machine, i.e., its read tape head cannot move to the left. This motivates defining the IDs of U as all tuples of {machine state, contents of output tape, contents of work tape(s), and contents of input tape *at or to the right of the input tape read head up to the end of the prefix codeword on the input tape*}.

In addition, I require that U can only halt if it has reinitialized its work tape(s). So all IDs with U in its halt state and output tape containing σ have the same (blank) work tape(s). This means that when U halts there is no "relic" recorded in the work tape(s) of what the original contents of the input tape was. In addition, by the precise definition above of IDs, all IDs with U in its halt state and output tape containing σ have no information concerning the contents of the input tape. So there is a unique ID with U in its halt state and output tape containing σ; it does not matter what input string to U was used to compute σ.

To simplify notation, let f be the transition function of U, i.e., write $\pi(v' \mid v) = \delta_{v', f(v)}$. Iterating f from a starting ID v_0 eventually results in an ID v' that specifies that U has halted. That v' is a fixed point of U. Write $\phi(v_0)$ for that fixed point arising from the ID v_0 (i.e., ϕ is the partial function computed by U). Also write $N(v_0)$ for the iteration at which U halts (with output value $\phi(v_0)$). Finally, define I^σ as the set of starting IDs that compute σ. Since we are interested in user distributions $\mathcal{P}_0$ that are guaranteed to compute σ, from now on I restrict attention to $\mathcal{P}_0$ whose support lies within I^σ.

There are several ways to define "the expected work for U to compute σ" using a calculator computer. In two of the most natural approaches, for any specific $v_0 \in I^\sigma$, the computer is run some number of iterations, after which v gets copied to the actuator and then reset, and the total amount of work is tallied. Where these approaches differ is in their rules for "when v gets copied to the actuator and reset".

In one approach we start with some specified $v_0 \in I^\sigma$ and then

1. Run the computer until it halts (with output σ) at timestep $N(v_0)$;

2. Copy that ending v (which is just σ) to the actuator;

3. Set v to its next value, v'_0, copied over from the sensor;

4. Cease to exist.

In this approach, an iteration of the calculator is identified as the sequence of iterations of f that takes v_0 to a halt state. So different v_0 will be identified with different numbers of iterations of f.

A second approach is the same as this first approach, except that we replace step (1) in this list with iterating f starting from the specified $v_0 \in I^\sigma$ a total of τ times. We then consider the limit as $\tau \to \infty$. Since v_0 is a starting state, we are guaranteed that under this limit, when we reach step (4) the computer has halted, and the value σ has been copied to the actuator. Moreover, as shown below, the minimal amount of work expended converges under this limit.

To evaluate the expected work in the first approach, combine the fact that π is a deterministic function, Eq. (9), and the restriction on the support of $\mathcal{P}_0$ to write

$$
\begin{aligned}
-\sum_{v\in I^\sigma} &\mathcal{P}_0(v)\ln[\mathcal{G}_0(v)] + \sum_{v\in I^\sigma}\mathcal{P}_{N(v)}(\phi(v))\ln\Big[\mathcal{G}_{N(v)}(\phi(v))\Big] \\
&= -\sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln[\mathcal{G}_0(v)] + \sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\mathcal{G}_{N(v)}(\phi(v))\Big] \\
&= -\sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln[\mathcal{G}_0(v)] + \sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\sum_{v' : f^{N(v)}(v')=\phi(v)}\mathcal{G}_0(v')\Big] \\
&= \sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\frac{\sum_{v' : f^{N(v)}(v')=\phi(v)}\mathcal{G}_0(v')}{\mathcal{G}_0(v)}\Big] \qquad (26)
\end{aligned}
$$

So the optimal $\mathcal{P}_0$ is a delta function about the $v \in I^\sigma$ that minimizes

$$
\frac{\sum_{v' : f^{N(v)}(v')=\phi(v)}\mathcal{G}_0(v')}{\mathcal{G}_0(v)} = \frac{\sum_{v' : f^{N(v)}(v')=\phi(v)} 2^{-\ell(v')}}{2^{-\ell(v)}} \qquad (27)
$$

and the associated minimal amount of work is

$$
\begin{aligned}
kT\ln(2)\mathrm{min}_{v\in I^\sigma}&\Big[\ell(v) + \log\Big(\mathcal{G}_0(\{v' : f^{N(v)}(v') = \phi(v)\})\Big) + \log\lambda\Big] \\
&= kT\ln(2)\mathrm{min}_{v\in I^\sigma}\Big[\ell(v) + \log\Big(\sum_{v' : f^{N(v)}(v')=\phi(v)} 2^{-\ell(v')}\Big)\Big] \qquad (28)
\end{aligned}
$$

where λ is the normalization constant for Chaitin's omega, i.e., the halting probability for U.

Intuitively, in this first approach, the amount of work for computing σ from some $v \in I^\sigma$ is given by the difference of two terms. The first is the length of v, i.e., how unlikely v is under $\mathcal{G}_0$. Or to put it another way, it is "how much information" there is in the initial ID of U. The second term is 'how much information" how much information there is concerning the initial ID of U by the time the computation ends. The bigger the drop in the amount of information concerning the initial ID, the more work is required to compute σ from v.

In contrast, in the second approach, the analogous analysis shows that the expected work is

$$
\begin{aligned}
-\sum_{v\in I^\sigma}&\mathcal{P}_0(v)\ln[\mathcal{G}_0(v)] + \lim_{\tau\to\infty}\Big\{\sum_{v\in I^\sigma}\mathcal{P}_\tau(\phi(v))\ln\Big[\mathcal{G}_\tau(\phi(v))\Big]\Big\} \\
&= -\sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln[\mathcal{G}_0(v)] + \lim_{\tau\to\infty}\Big\{\sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\sum_{v' : f^\tau(v')=\phi(v)}\mathcal{G}_0(v')\Big]\Big\} \\
&= \sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\frac{\lim_{\tau\to\infty}\sum_{v'\in I^\sigma : N(v')\leq\tau}\mathcal{G}_0(v')}{\mathcal{G}_0(v)}\Big] \\
&= \sum_{v\in I^\sigma}\mathcal{P}_0(v)\ln\Big[\frac{\mathcal{G}_0(I^\sigma)}{\mathcal{G}_0(v)}\Big] \qquad (29)
\end{aligned}
$$

where the penultimate step uses the fact that $\phi(v)$ is a fixed point for all $v \in I^\sigma$, and the last step uses the fact that all $v \in I^\sigma$ eventually halt.

So defining expected work using this second approach, the optimal $\mathcal{P}_0$ is a delta function about the $v \in I^\sigma$ that minimizes $\mathscr{G}_0(v) \propto 2^{-\ell(v)}$. But that is just the v_0 of minimal length in the set of all v_0 that result in output σ. The associated minimal expected work is

$$kT \ln(2)\Big(K_U(\sigma) + \log[\mathscr{G}_0(I^\sigma)] + \log \lambda\Big) \tag{30}$$

as claimed in Eq. (10).